B2

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 19/00 — A2

(11) International Publication Number: WO 00/41122

(43) International Publication Date: 13 July 2000 (13.07.00)

(21) International Application Number: PCT/US00/00167

(22) International Filing Date: 5 January 2000 (05.01.00)

(30) Priority Data:

60/114,806	5 January 1999 (05.01.99)	US
Not furnished	4 January 2000 (04.01.00)	US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Applications

US	60/114,806 (CIP)
Filed on	5 January 1999 (05.01.99)
US	Not furnished (CIP)
Filed on	4 January 2000 (04.01.00)

(71) Applicant *(for all designated States except US)*: CURAGEN CORPORATION [US/US]; 11th floor, 555 Long Wharf Drive, New Haven, CT 06511 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: BADER, Joel, S. [US/US]; 36 Ogden Road, Stamford, CT 06903 (US). LIU, Yi [CN/US]; 470 Prospect Street, #53, New Haven, CT 06511 (US). GOLD, Stephen [US/US]; 36 Whitting Farm Road, Branford, CT 06405 (US). DZIUDA, Darius [PL/US]; 1298 Hartford Turnpike, 9E, North Haven, CT 06473 (US). GUSEV, Vladimir [UA/US]; 1209 Durham Road, Madison, CT 06443 (US). JUDSON, Richard, S. [US/US]; 42 Barker Hill Drive, Guilford, CT 06437 (US). WENT, Gregory, T. [US/US]; 34 Scotland Avenue, Madison, CT 06443 (US).

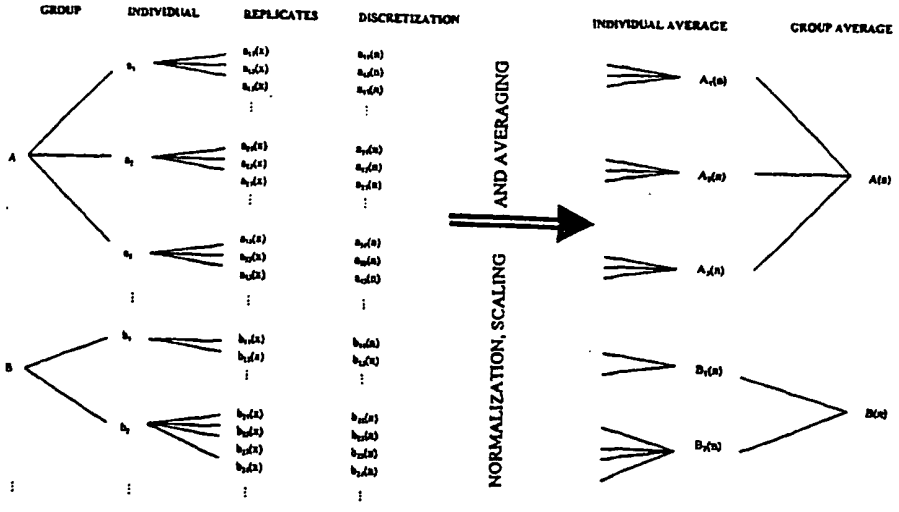
(74) Agent: ELRIFI, Ivor, R.; Mintz, Levin, Cohn, Ferris, Glovsky, and Popeo, P.C., One Financial Center, Boston, MA 02111 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published

Without international search report and to be republished upon receipt of that report.

(54) Title: NORMALIZATION, SCALING, AND DIFFERENCE FINDING AMONG DATA SETS



(57) Abstract

This invention relates to a method of identifying a difference between at least two data sets made up of ordered elements utilizing internal features within the data sets for calculations relating to normalization, scaling, and difference finding.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

NORMALIZATION, SCALING, AND DIFFERENCE FINDING AMONG DATA SETS

FIELD OF THE INVENTION

This invention relates to statistical analysis of differences between at least two data sets.

RELATED APPLICATIONS

The present patent application claims priority to the United States provisional patent application U.S.S.N. 60/114,806, entitled "Scaling and Normalization" filed January 5, 1999, which is incorporated herein by reference in its entirety.

BACKGROUND OF THE INVENTION

Measurements of the expression levels of individual genes within a cell provide a wealth of information about cellular processes. This is done by extracting messenger RNA molecules (mRNA) from a cell, possibly converting these to more stable cDNA molecules, and measuring the concentrations of each individual species by methods such as differential display or hybridization. A typical analysis strategy is to identify genes whose expression levels differ between particular biological states. One difficulty in performing such an analysis is that experimental measurements of expression levels include variation due to noise. Distinguishing the true differences from the false differences (those due simply to noise) has presented a challenge for gene expression analysis. The relevance of this problem is that genes that are differentially regulated can be converted to commercial products, including protein therapeutics, antibody targets, therapeutic markers, as well as conventional drug targets.

More broadly, similar data sets may arise in any of a number of ways. Any experiment or study in which one group is compared with another may provide data sets whose elements include the effects of noise. Noise and other uncontrolled variations within and among data sets arising in such experiments make comparisons between the data sets more difficult, and present challenges in evaluating the results appropriately.

One of the major sources of noise in such experiments, including gene expression experiments, is that the amount of material analyzed, such as mRNA or cDNA, can differ from experiment to experiment, or among the replicates of a single experiment. Data analysis

strategies typically account for this overall variation by performing a global scaling of all the measurements from such experiments. For example, if sample A has twice the overall cDNA concentration of sample B, then the expression level for a gene in sample B must be doubled before comparison with sample A.

Often, however, such an overall scaling is not sufficient to discriminate between true differences and those that can be attributed to noise. One source of difficulty is identifying the particular features in the data set or sets that can be used as scaling landmarks. It is not always possible to identify *a priori* such unchanging features ahead of time

An additional source of noise generally present in experimental studies is noise from analytical instruments and methods. In differential display experiments, for example, the amount of gene expression is related to the amount of PCR product generated in an amplification reaction. The amount of product can depend on the activity of the polymerase enzymes as well as the length of a fragment being replicated. If the enzyme functions effectively, the amount of PCR product is uniformly high from small fragments to long fragments. If the enzyme activity is less effective, however, the amount of PCR product can be relatively less for long fragments than for short fragments. An overall scaling does not account for the non-uniform tapering of the signal with the size of the amplicon.

There thus remains a strong need for counteracting and overcoming the effects of noise in comparing data sets in an experimental study. There is a lack of adequate means for identifying constant or unvarying components in a data set, which may serve as reference markers in normalizing, scaling and distinguishing differences among data sets in such a study. There further is a significant need for a means to identify scaling landmarks automatically in data sets being compared to one another. In a particular framework addressed in this invention, there is a need for robust methods that normalize scale and find differences in experiments related to the differential expression of genes in cells and tissues subjected to specific experimental treatments. These and comparable needs are addressed by the present invention.

SUMMARY OF THE INVENTION

The present invention discloses a method of identifying a difference between at least two groups, wherein each group comprises a data set containing ordered elements. The method includes the steps of: (a) providing a first group having one or more elements in a first data set;

(b) applying at least one transformation to said first data set to provide a transformed data set, wherein said transformation is a calculation selected from a normalizing calculation, an averaging calculation and a scaling calculation; and (c) distinguishing differences, if present, between elements of said first transformed data set and a second groups having one or more elements in a second data set; thereby identifying a difference between the data sets.

In some embodiments, the method corrects the effects of the noise prior to distinguishing the differences. Prior to normalization, averaging, and/or scaling calculations, selected regions in a data set that do not contain useful data may be masked, and regions that have a higher information content may be highlighted. These include regions where the signal intensity in the data set is either too low (noise) or too high (saturation) for accurate measurement, or is at locations of local peaks. In one embodiment, the noise includes low frequency noise. In another embodiment, the noise includes jiggle. In the latter embodiment, the jiggle includes positional shifts of elements between different data sets and signal alignment within a data set. In a further embodiment, the jiggle is corrected. In another embodiment, correction of jiggle may be considered as signal alignment between two or more data sets.

In some embodiments, the elements of a data set may represent, for example, a trace, such as a trace arising in an electrophoretogram or a chromatogram. In an alternative embodiment, an element of a data set represents a position in a reagent array. In a further significant embodiment the position in the reagent array determines one extent of matching between a reagent affixed to the array at the position and a sample contacting the array position. For example, the reagent may be a first nucleic acid and the sample may include a second nucleic acid.

In a further embodiment, a data set is obtained in an experiment related to identifying significant differences in gene expression. In some embodiments, a group includes more than one individual. Additionally, a data set of the method may be subjected to a masking operation. In further embodiments, the data set for each group is obtained by applying at least one calculation, chosen from a normalizing calculation, an averaging calculation and a scaling calculation, to the data sets from each individual. In still further embodiments, each individual provides at least one replicate sample that is employed to provide a trace. In such embodiments, the traces for each individual are advantageously transformed by applying at least one of a normalizing calculation, an averaging calculation and a scaling calculation to the trace from each

replicate; and in additional such embodiments, the traces for each replicate are discretized prior to the normalizing, the averaging and/or the scaling.

In another embodiment, the normalization includes adjusting each data set such that a subset of elements in each data set has similar or identical values.

In further embodiments, the averaging includes calculating an average for a location or for a discretized position across a collection of data sets. The average may be an unweighted average or a weighted average.

In additional embodiments, the scaling includes a calculation that causes a first data set to resemble a second data set except that an element in the scaled first data set whose intensity differs significantly from the intensity of the element in the second data set at the same location or the same position contributes to identifying the difference between the data sets. In further embodiments, the scaling includes calculating a distance between the data sets, or calculating a similarity between the data sets. In particularly embodiments, the scaling calculation employs a scaling function; and in other embodiments, the scaling function is a basis set expansion, such as a piecewise linear basis set or a direct product of basis functions.

In yet additional embodiments, successive iterations of a cycle that includes at least one of a normalization calculation, an averaging calculation and a scaling calculation are carried out until a specified termination condition has been satisfied. In particularly embodiments, the termination condition is that the transformed data set has converged. Alternatively, the termination condition is that a predetermined number of iterations has been reached.

In a further embodiment, the distinguishing of differences among the elements of the transformed data sets includes application of a difference finding algorithm.

The invention also discloses a display means that displays a representation of a difference between data sets, and also discloses the representation itself, wherein the representation is obtained in general by applying methods disclosed herein to the data sets.

Brief Description Of The Drawing

FIG. 1 is a graphic representation of jiggle arising between two traces.

FIG. 2 is a schematic diagram illustrating the flow from different groups to the transformed data sets for those groups.

FIG. 3 is a schematic estimation of σ, the experimental noise in a data set.

FIG. 4. is a schematic representation of averages of three replicate raw traces for each individual animal in Example 1, prior to normalization or scaling.

FIG. 5. is a schematic representation of averages of 3 normalized replicate traces for each individual animal in Example 1.

FIG. 6. is a schematic representation of scaling factors employed to scale the phenobarbital-treated individual average to the sterile-water-treated individual average in Example 1, obtained as a result of iterative scaling.

FIG. 7. is a schematic representation of iteratively scaled traces for each individual animal in Example 1.

DETAILED DESCRIPTION OF THE INVENTION

The invention discloses methods for normalizing, scaling, and difference finding that may be used in any experimental study, including gene expression data, in which noise and other uncontrolled variations exist between data sets. In particular embodiments, these methods have been adapted for differential display experiments, in which gene expression levels are represented by fragment intensities in, for example, an electrophoresis trace. The methods are also applicable to, for example, hybridization experiments, such as those employed with nucleic acid microchip arrays, as well as other experiments not related to gene expression.

The present invention discloses a method of identifying a difference between at least two groups, wherein each group comprises a data set containing ordered elements. The method includes the steps of: (a) providing a first group having one or more elements in a first data set; (b) applying at least one transformation to said first data set to provide a transformed data set, wherein said transformation is a calculation selected from a normalizing calculation, an averaging calculation and a scaling calculation; and (c) distinguishing differences, if present, between elements of said first transformed data set and a second groups having one or more elements in a second data set; thereby identifying a difference between the data sets.

Each data set can be represented as a set of discretized intensity values, or elements in the data set. The intensity of an element may include effects of noise, as described below, and the method operates to correct the differences for the effects of the noise. In one embodiment, the

noise includes low frequency noise. In another embodiment, the noise includes differences in jiggle. In the latter embodiment, the jiggle includes positional phase shifts of elements between different data sets. The invention also discloses a display means that displays a representation of a difference between data sets, and also discloses the representation itself, wherein the representation is obtained in general by applying methods disclosed herein to the data sets.

As used herein, "representation" relates to any graphical, visual, or equivalent non-verbal display that provides an image of the results, such as differences between data sets, obtained according to the methods of the present invention. More specifically, a "representation" of the invention is obtained by transforming the quantitative results gathered by experiments underlying the invention. Examples of such data include, by way of non-limiting example, traces from differential gene expression, and intensities from an array, and/or equivalent types of experimental parameter.

In some embodiments, a representation of the invention is generated by algorithms executed in a computer and is suitable for display on a display means, such as a display screen or monitor, employed in the operation of the computer. The representation is also suitable for storing in a storage module or data archive of such a computer. It is still further suitable for printing from the computer onto a medium such as paper or equivalent physical medium, and for recording it onto a portable storage medium, including, for example, magnetic media, CD ROMs and equivalent storage media. As used herein, "display means" includes any of the objects and media identified above in this paragraph, as well as equivalent apparatuses and objects suitable for displaying the results of computational processes for visual inspection.

In addition, "normalization" is defined herein as a means for standardizing or correcting elements in a data set, for example, but not by way of limitation, for correcting overall signal strength within a given data set. Features of given elements to be normalized are first identified within a data set. For example, one such feature may be the median peak height of signals within a data set. A summary statistic for the given feature is generated for a data set, and used to normalize the elements, as described below, to allow comparisons across data sets. Algorithms that are designed to either mask or highlight chosen features identified among the elements of a data set may be applied prior to normalization, averaging or scaling. Such features include low intensity signal regions that comprise noise, high intensity signal regions that comprise saturation zones, and local maxima that comprise peaks. "Averaging" is defined as combining multiple

data sets to generate one average representative data set. Averages are combined into the representative data set in such a way that noise from any one data set so combined does not affect any other data sets used to generate the average. "Scaling" is defined as a correction for low frequency difference is signal strength across data sets.

The data sets may arise in any of a number of ways. Any experiment or study in which one group is compared with another may provide the data sets employed in the invention. Such groups may be distinguished by the experimental conditions experienced by the respective groups, or by the experimental state characterizing the respective groups. Experimental subjects may be animate or inanimate, or may be inanimate samples derived from animate subjects.

In some embodiments, display means and representations, the data sets arise from experiments conducted in investigations in which identification of the differential expression of a gene or genes between data sets from at least one experimental group and at least one control group is sought. In certain embodiments of the invention, such differential expression arises in GeneCalling™ experiments. See, *e.g.*, United States Patent No. 5,871,697; Shimkets *et al.*, *Nat. Biotechnol.* 17: 798-803 (1999). In other embodiments of the invention, such differential expression is evaluated using nucleic acid microchip arrays in order to detect the presence, absence or extent of expression of a gene or gene fragment. Any alternative, equivalent differential expression formats and methods of analysis are encompassed within the scope of the present invention as well.

Various types of noise may arise during the course of gathering the data elements comprising the data sets. Nonlimiting examples of noise include intensity noise and extension noise leading to longitudinal differences. Intensity noise includes relatively high frequency noise such as that commonly associated with short-time fluctuations in the electronic and/or mechanical components of an experimental system. High frequency noise may be defined as having a frequency greater than about 1 Hz. Examples of high frequency noise include shot noise in photodetectors and comparable electronic noise arising in the various electronic components and circuits of an experimental instrument employed in gathering the data elements of a data set.

The methods of the present invention can minimize or eliminate low frequency noise. Low frequency noise has a frequency less than about 1 Hz, and may have frequencies less than about 0.1 Hz, or less than about 0.01 Hz, or even less than about 0.001 Hz or lower. Such low

frequency noise may arise during an experiment, for example, by decay of activity of a reagent, catalyst or enzyme during the course of preparing a sample that is applied to generate a data set. Alternatively, an uncompensated low frequency change in response of an electronic instrument may arise during the time in which a data set is being gathered. Additionally, if an array is being used to generate the data set, uncompensated variations in detection across the various positions and/or dimensions of the array may arise that behave as low frequency noise (i.e., they may be considered as low frequency noise even though an array may be subjected to simultaneous detection of all the sample points on the array, since positional variations behave as if they have a long wavelength across the array.) Equivalent sources of low frequency noise are also encompassed in this definition. Normalization and scaling algorithms employed are particularly effective in minimizing or eliminating the effects of low frequency noise.

An additional detrimental effect that may arise in identifying differences between data sets is termed "jiggle". By this term is meant that the elements of one data set are offset in a longitudinal direction in comparison with the elements of a second data set with which the first data set is being compared. Longitudinal displacement relates to variation in the location or discretized position of a particular feature in a trace even though the feature appears in the traces of more than one group. By way of nonlimiting example, uncompensated variation in the location or discretized position of the feature may occur due to variations in physical or chemical conditions during the process of accumulating the data elements of the various data sets being considered. Such a variation, or jiggle, may be considered to be low frequency noise in the longitudinal, or positional, direction. Jiggle is illustrated in FIG. 1. In this figure, two discretized, normalized data sets, A(n) and B(n), are shown. A(n) and B(n) should be thought of as each representing the same feature. Nevertheless, they are displayed with a jiggle of 1.75 units on the *n* axis. It is an additional aspect of the present invention that the normalization and scaling algorithms employed are particularly effective in compensating for and/or overcoming the effects of low frequency longitudinal noise. Such procedures, as employed in the methods of the present invention, largely or completely eliminate the jiggle and, referring to FIG. 1, restore the overlap of the points for A(n) and B(n). Compensation for jiggle as shown in FIG. 1 is also termed "signal alignment."

Groups, Individuals, Replicates and Transformed Data

A hierarchy of notation is used herein to indicate data elements and/or the data sets. These notations are discussed below and furthermore are illustrated in the flow diagram presented in FIG. 2. Raw, *i.e.* untreated or untransformed, data arise from the carrying out the experiments on actual samples obtained from experimental groups. A "group" represents a particular experimental state or condition. The groups are denoted herein in capital letters A, B, ... without any indices or delimiters. As shown in FIG. 2, at least two groups comprise the subject matter on which the methods, display means and representations of the present invention are based.

Each group gives rise to data elements comprising data sets after samples from the groups have been subjected to a given experimental method of detection or analysis. Experimental data not transformed by any calculations of the methods disclosed herein are designated using lower case letters together with at least one index or delimiter *i*, shown, for example, by a_i, b_i, ... (see FIG. 2). A group may be initially composed of one or more individuals. The number of individuals is not fixed or constant, but may vary. Each individual in the group is subjected to the same experimental conditions or experimental state. For the case of animate groups, each individual may represent an individual animal, a plant (such as a seedling), or a set of cells grown in cell or tissue culture. Correspondingly, for inanimate groups, each individual may represent, again by way of nonlimiting example, a separate execution of a particular experimental protocol such as a synthetic or preparative procedure, or the implementation of a particular set of physical conditions on separate samples or objects. Equivalent ways of designating individuals of a group are encompassed within the scope of the present invention. In general, the data sets obtained from the individuals of a group may be transformed by any one or more of the normalization, averaging and scaling calculations of this invention in arriving at the differences determined by the present methods.

As a further hierarchical subclassification, each individual of a group may furnish one or more replicate samples for detection or analysis according to the experimental method employed. Such replicates also represent raw, or untreated, data. Each replicate of an individual is designated with a second index or delimiter *j*, shown, for example, by a_{ij}, b_{ij}, ... (see FIG. 2). As shown for illustration in FIG. 2, the number of replicates may vary due to experimental circumstances. Commonly replicates are obtained by repetitive sampling from the same

individual. In general, the data sets obtained from the replicates of a particular individual may be operated upon by any one or more of the normalization, averaging and scaling calculations of the present invention in arriving at the differences determined by the present methods. The normalization, averaging and/or scaling calculations that are applied to the replicates may be applied prior to, or simultaneously with, the similar calculations applied to the individuals and discussed in the preceding paragraph.

In many of the detection or analytical methods employed in the experiments underlying the gathering of the presently disclosed data sets, continuous traces of an experimental intensity as a function of a longitudinal variable such as time, elution volume or distance are obtained. Such traces arise, for example, in the use of chromatographic or electrophoretic methods of detection or analysis. Since the traces are continuous, each data set may be considered to be comprised of an infinite number of data elements designated using a further delimiter, $a_{ij}(x)$, where x denotes the continuous longitudinal dimension of the analytical method. (It may be noted that use of alternative analytical or detection methods, for example use of arrays with discrete positions on them, does not generate a continuous trace. Such data sets, therefore, in general need not carry the additional delimiter x.) It is convenient for virtually all calculations carried out as disclosed herein, using computers with discrete memory locations for separate data elements, to discretize a continuous trace into discrete intensities at specified locations, or discrete positions n, on the trace. As used herein, the delimiter n replaces the delimiter x when the intensity trace has been discretized; i.e., $a_{ij}(x)$ becomes $a_{ij}(n)$ (see FIG. 2).

As used herein, any data sets that have been transformed using the calculations disclosed herein are designated in upper case letters including an index and/or a delimiter. As noted, the transformations may include at least one operation chosen from among normalization, averaging and scaling. A transformation that operates to combine the replicates of an individual while leaving the individuals of a group intact results in a transformed data set indicated by one index and a delimiter, $A_i(n)$, $B_i(n)$, ... (see FIG. 2). Further transformation that operates to combine the individuals of a group to provide a single data set for an entire group is designated by a delimiter only, as shown, for example, by A(n), B(n), ... (see FIG. 2). Conversely, in experimental methods of analysis or detection that do not rely on developing traces, the $A_i(n)$, $B_i(n)$, ... are obtained directly without discretization. They may still arise from replicate samples, however.

For example, if the detection method is based on an array, one or more positions in the array may represent the results of one or more replicates, respectively.

A particular embodiment of a data set envisioned in the present invention is differential display. In a differential display experiment involving gene expression, mRNA is extracted from sample, converted to cDNA, and digested with restriction enzymes into fragments (United States Patent No. 5,871,697; Shimkets *et al.*, *Nat. Biotechnol.* 17: 798-803 (1999)). The fragments are then separated according to length using electrophoresis. Although nucleic acids consist of an integer number of nucleotides, their electrophoretic transport properties also depend on the nucleotide composition. Electrophoresis experiments currently in use measure the length of a nucleic acid fragment determined electrophoretically to a precision of 0.1 nt, and the actual value of the electrophoretic length is usually within 1-2 nt of the actual number of nucleotides in the fragment. The intensity signal a(x) of the electrophoresis trace for sample A represents the amount of fragments of electrophoretic length x generated from the sample. Of course, the intensity a(x) also depends on the particular restriction enzymes used to generate fragments; for simplicity, this dependence is suppressed in the notation. Sometimes the intensity at length x corresponds to a single fragment; sometimes multiple fragments have the same length and their signals are combined; sometimes no fragments are present and a(x) is a baseline signal. Because a(x) is a measured intensity, it should be a positive quantity. Mathematical operations used during signal processing, such as the subtraction of a baseline, might result in negative values at certain locations a(x). If negative values exist, their magnitude should be preferably of the same order as the measurement error in the data set.

To detect differences, the intensity a(x) from samples in group A is compared with the intensity b(x) generated using an identical protocol from samples in group B. Some inadvertent differences between A and B can be attributed to underlying genetic variation in the individuals chosen. For example, samples A and B may include organisms or individuals having an allelic variation between them that generates a difference in the measured expression levels but has no biological relevance in the context of the particular experimental study. In differential display, for example, a neutral single nucleotide polymorphism (SNP) can add or remove a band. For this reason it is preferable to include multiple organisms or individuals for samples A and B to control for these types of individual differences. In general, as noted earlier, the expression

profile of the i^{th} individual of group A is denoted $a_i(x)$, and similarly $b_i(x)$ is the expression profile for individual i of group B.

Furthermore, it is preferable to have multiple experimental replicates of the expression profiles for each organism or individual. The j^{th} expression profile, or replicate, of the i^{th} individual of group A is denoted as $a_{ij}(x)$, and similarly for group B. Thus, each group may have one or more individuals, and each individual may have one or more replicates. More elaborate hierarchies are also possible and may be analyzed directly with the methods outlined below.

An alternative embodiment of an experimental system relates to hybridization. In this embodiment, let $a_{ij}(x)$ represent the intensity from the j^{th} experimental replicate of the i^{th} organism in group A measured at position x on a hybridization array or chip. Here x is a two-dimensional coordinate that identifies the location of a particular spot on the hybridization surface. The term trace is used herein to denote a one-dimensional data set, and the term array or hybridization data is used herein to represent a two-dimensional data set. Terms such as data set, signal, and intensity may represent one- or two-dimensional data sets. Furthermore, repeated experiments, such as hybridization experiments conducted on a series of biological samples collected over time, may have additional dimensions. By way of nonlimiting example, each time point in a time course study generates a two-dimensional plane of data, and the time coordinate adds a third dimension. The methods disclosed herein are applicable to data sets such as these, as well as to those of the preceding paragraphs. In full generality the methods disclosed herein are generally applicable to any experimental study that generates multi-dimensional data sets. In particular cases, attention may be restricted to a particular dimensionality of data sets, as the specific character of the study may provide. Furthermore, as used herein the terms "signal" and "intensity" may be considered interchangeable references to either measured, normalized, scaled, or averaged data.

Data sets can have many representations in the memory of a computer. Here it is assumed that each data set can be represented as a set of discretized intensity values, or elements in the data set. Although it is not necessary to use a regular grid to store the intensity, it is convenient to do so. Using a regular grid in one dimension, an intensity a(x) is stored at locations $x = 0, \Delta x, 2\Delta x, \ldots, l\Delta x$ (where $l\Delta x = L$, and L represents the full length of a trace in the longitudinal direction). In two dimensions, a(x) is stored at locations (x_1,x_2) where $x_1 = 0, \Delta x, 2\Delta x, \ldots, l\Delta x$, and $x_2 = 0, \Delta y, 2\Delta y, \ldots, m\Delta y$ ($m\Delta y = M$). In d-dimensions, a(x) is stored at

locations ($x_1, x_2, \ldots, x_d$) where, for $k = 1, 2, \ldots, d$, x_k takes on the values $0, \Delta x_k, 2\Delta x_k, \ldots, l_k\Delta x_k$. For convenience, we also introduce the notation a(n) where n is a d-tuple of integers ($n_1, n_2, \ldots, n_d$) and a(n) corresponds to the intensity a(x) where $x_k = n_k\Delta x_k$.

For electrophoresis data, such as that generated by differential-display data, Δx is preferably close to the reproducibility of the instrument. With currently available instruments, a value of 0.1 nt is preferable. For raw hybridization images, a value corresponding to a single pixel in an image is preferable. For processed hybridization images, it is preferable that each discretization point represents an individual spot with a different probe.

Methods of Calculation, Algorithms

The invention allows for the identification of a difference between data sets, wherein each data set contains data elements as described above. The differences are identified by operating on at least two transformed data sets, A(n) and B(n), to discern particular discrete positions *n* at which differences that exceed a lower limit of distinction are found. The methods are that they use algorithms that automatically identify scaling landmarks that may exist in the data sets being evaluated.

Details of the ways of identifying statistically significant differences are presented in the following sections and in the EXAMPLE.

Masks for Noise, Saturation, and Peaks

It is useful to mask out regions in the data set that do not contain useful data and to highlight regions that have a higher information content. These include regions in which the intensity is too low or too high for an accurate measurement and locations of peaks. A series of masks m(n) records this information for a data set a(n).

The noise mask $m_{noise}(n)$ depends on a noise level I_{noise} that characterized the experimental uncertainty in the measured intensity. This uncertainty may be estimated, for example, as the standard deviation of the background signal obtained for a blank or control sample. I_{noise} may also be preferably assigned a value that is a small multiple (0.2X to 5X) of the low end of the dynamic range of the detection instrument. The noise mask is calculated as follows:

- For each position n

- $m_{noise}(n) = 1$ if $a(n) < I_{noise}$
- $m_{noise}(n) = 0$ otherwise.

The saturation mask $m_{sat}(n)$ marks data collected near the upper limit of the detection range of an instrument. The mask depends on a saturation level, as follows:

- For each position n
  - $m_{sat}(n) = 1$ if $a(n) > I_{sat}$.
  - $m_{sat}(n) = 0$ otherwise.

The threshold I_{sat} is preferably close to the high end of the dynamic range of the detection instrument (0.95X or to 1X).

Preferably, points that are marked as not saturated are checked for saturation in a second pass that depends on a saturation width w_{sat} as follows:

- For each position n
  - $m_{sat}(n) = 1$ if a(n) is part of a plateau of constant value over a range of $\pm w_{sat}$ in each dimension. Thus, in one dimension, if $a(n) = a(n+1) = a(n+2) = \ldots = a(n+w_{sat})$ and $a(n) = a(n-1) = a(n-2) = \ldots = a(n-w_{sat})$, then $m_{sat}(n) = 1$ for each of these points. Less preferably, $m_{sat}(n) = 1$ only for the center point; the remaining points require their own saturation checks.
  - $m_{sat}(n) = 0$ otherwise.

For differential display data, $w_{sat}\Delta x = 0.2$ nt is preferable, and I_{sat} corresponds to the camera intensity at saturation.

Next points that are local maxima are identified as peaks. A parameter w_{peak} defines the half-width for a peak as follows:

- For each position n
  - In multi-dimensions, $m_{peak}(n) = 1$ if $a(n + \Delta n') < a(n + \Delta n)$ for all Δn and $\Delta n'$ such that $\Delta n'$ is farther than Δn from n, $\Delta n'$ is no farther than w_{peak} from n, and Δn and $\Delta n'$ are identical except for a single dimension in which they differ by 1. For this purpose, distances may be calculated by any of a number of methods,

including the Euclidean metric, the Manhattan metric, or the maximum absolute difference in any dimension. In one dimension, $m_{peak}(n) = 1$ if $a(n-w_{peak}) < \ldots < a(n-2) < a(n-1) < a(n) > a(n+1) > a(n+2) > \ldots > a(n+w_{peak})$. Preferably, $a(n+w_{peak}) > I_{noise}$ and $a(n-w_{peak}) > I_{noise}$ as well.

- $m_{peak}(n) = 0$ otherwise.

For one-dimensional differential display data, $w_{peak}\Delta x = 0.3$ nt is preferable. For hybridization, if an image has already been processed such that each point represents a different probe on the surface, then each point is regarded as a peak.

Normalization

A data set is normalized by first determining the peak intensity values a(n) at locations where the peak mask $m_{peak}(n) = 1$. A summary peak intensity I_{peak} is calculated from the individual values. If desired, the values can be rank-ordered, and I_{peak} can be defined as the 75th percentile value (75% of peaks are smaller in value; 25% of peaks are larger). Other methods include using a different percentile, for example the median, or calculating an average value. Rank-order selection methods are more robust than averages.

After a peak intensity has been calculated, the data set is rescaled by multiplying each point a(n) by the factor (I_{norm}/I_{peak}), where I_{norm} is identical for each data set and sets a convenient scale. Although the precise choice for I_{norm} is arbitrary, a value such as 100 is convenient.

The noise threshold I_{noise} may also be subject to the same normalization. Alternatively, I_{noise} may be set to a fixed value. A preferable fixed value for differential display data is $I_{noise} = 10$.

Averaging

Averaging is an operation that is applied to a collection of data sets. The average A(n) of a collection of r data sets $a_1(n), a_2(n), \ldots, a_r(n)$ is calculated as

$$A(n) = \Sigma(i=1..r)\ w_i\ a_i(n) / [\ \Sigma(i=1..r)\ w_i\ ]$$

where w_i is a weighting applied to data set i. If $\Sigma(i=1 \ldots r)\ w_i = 0$, then another method must be used. It is preferable to use local information from the closest points where the summed weight does not vanish to estimate A(n). Most preferably, the value A(n) can be set equal to the

value at the closest point where the weight does not vanish. Alternatively, the unweighted values $a_i(n)$ may be used.

Any of a variety of weighting functions may be used, and may account for characteristics of a data set such as those discussed in the following. A preferred weighting function is $w_i = [1 - m_{sat,i}(n)]$ where $m_{sat,i}(n)$ is the saturation mask for data set i. Weights may also incorporate error estimates from the data sets. For example, suppose that the data set $a_i(n)$ is known with statistical error $e_i(n)$. Then a maximum likelihood estimate for A(n) is obtained by minimizing a chi-square statistic

$$\chi^2 = \Sigma(i=1..r) [A(n) - a_i(n)]^2 / e_i(n)^2$$

with respect to the final average A(n) to obtain $w_i = 1/e_i(n)^2$, or, if desired, $w_i = [1-m_{sat,i}(n)]/e_i(n)^2$. If $a_i(n)$ is itself derived from an average of other data sets, then the standard error of the mean is an appropriate choice for $e_i(n)$. If $a_i(n)$ is an unnormalized data set, then an appropriate choice for $e_i(n)$ is the background noise level I_{noise} defined previously. If $a_i(n)$ is a normalized data set, then it is appropriate to scale the noise level as well and use I_{noise}/I_{peak} for $e_i(n)$.

It is preferable to calculate a standard deviation $SD_A(n)$ to describe the distribution of data points leading to the average A(n). A preferable formula for $SD_A(n)$ is

$$SD(n) = [\Sigma(i=1..r) [A(n) - a_i(n)]^2 / (r-1)]^{1/2}.$$

The standard error for the average is preferably calculated as

$$E(n) = SD(n)/r^{1/2}.$$

Similarity and Difference

In preparation for describing the scaling operation below, it is necessary to determine the extent of agreement between two data sets. This extent of agreement can be measured, by way of nonlimiting example, by the distance Dist[A,B] or the similarity Sim[A,B] between two data sets A(n) and B(n).

Two possible formulas for the difference Dist[A,B] are

$$Dist[A,B] = \Sigma_n w[A(n),B(n)] dist[A(n),B(n)] \text{ and}$$

$$Dist[A,B] = \Sigma_n w[A(n),B(n)] dist[A(n),B(n)] / \Sigma_n w[A(n),B(n)].$$

The term dist[A(n),B(n)] is a function that measures the distance between two values A(n) and B(n). The term w[A(n),B(n)] is a mask that determines whether the data points at location n should be included in the calculation. The second formula is preferable.

A preferable formula for the distance dist(a,b) for two numbers a and b is

dist(a,b) = $[\ln(a/b)]^2$, where ln() is the natural logarithm. Here a and b must be regularized to prevent values close to 0 from causing a divergence. This can be accomplished, for example, by replacing a or b by a minimum value I_{min} if either is smaller than I_{min}, or by adding a positive constant to raise all values A(n) and B(n) above 0.

Other acceptable formulas are as follows:

the absolute difference, dist(a,b) = | a - b | ;

the Euclidean distance, dist(a,b) = $[\ (a - b)^2\ ]^{1/2}$;

the square distance, dist(a,b) = $(a-b)^2$; or

any non-negative function F(a,b) that is 0 when a = b and increases with increasing |a-b| or increasing |ln(a/b)|.

A preferable formula for the weight w[A(n),B(n)] is

$w[A(n),B(n)] = w_{A(n)}w_{B(n)}$ if $dist'(a,b) < D_{max}$ and

w[A(n),B(n)] = 0 otherwise,

where $w_{A(n)}$ and $w_{B(n)}$ are weights for the individual data sets, dist'(a,b) is a distance measure and D_{max} is some maximum distance. Possible choices for the distance measure dist'(a,b) are the same as the choices for dist, but the same distance measure need not be used for both. A preferable choice is dist'(a,b) = |ln(a/b)| and $D_{max} = 3$.

The weight $w_{A(n)}$ is preferably $[1 - m_{A,noise}(n)][1 - m_{A,sat}(n)]$, and similarly for $w_{B(n)}$. Other acceptable alternatives are to use either the noise mask or the saturation mask, or to use no mask and set $w_{A(n)} = w_{B(n)} = 1$.

It is also acceptable to use w[A(n),B(n)] = 1.

A similarity Sim[A,B] between two data sets A and B may be defined as

$Sim[A,B] = \Sigma_n\ sim[A(n),B(n)]$

where the similarity sim(a,b) between two numbers a and b is larger when the quantities are larger and also larger when a and b are closer in value. A preferred method for calculating sim(a,b) is as follows. Plot the point (a,b) and measure the length r of its projection onto the line a=b and its distance d from the same line, as shown in the figure below. Then define

$$sim(a,b) = \rho \exp(-d^2/2\sigma^2)/[2\pi\sigma^2]^{1/2} \text{ where}$$

$$\rho = |a + b|/\sqrt{2},$$

$$d = |a - b|/\sqrt{2}, \text{ and}$$

σ characterizes the experimental noise in the data sets.

As depicted in FIG. 3, one method to estimate an appropriate value for σ is to calculate the slope m of the best linear regression line b = m a, then calculate σ as the root mean square residual of the points (A(n),B(n)) from the line b = ma,

$$\sigma = [\ \Sigma_n 2\ [\ (mA(n) - B(n)) / (m+1)\ ]^2 / (r-1)\ ]^{1/2}$$

where r-1 is the number of degrees of freedom in the fit.

Other preferable formulas for sim(a,b) are

$$sim(a,b) = F(\rho)\ G(d)$$

where $F(\rho)$ is an increasing function of ρ and G(d) is a decreasing function of d.

This algorithm is related to one of the literature as a method of decomposing spectra of multicomponent mixtures into separate spectra for each of the pure components.

Equivalent procedures for evaluating the similarity and difference between data sets is encompassed within the scope of the present invention.

Scaling

Scaling is an operation that is applied to a subordinate data set a(n) to bring it in closer agreement with a master data set A(n). A scaling algorithm optimizes a scaling function s(n) to minimize the distance or maximize the similarity between the scaled slave data set s(n)a(n) and the master data set A(n).

The scaling function s(n) may have various mathematical representations. One representation is a basis set expansion

$$s(n) = \Sigma_p c_p \phi_p(n)$$

where c_p is an expansion coefficient, $\phi_p(n)$ is the value of the p^{th} basis function at position n, and p ranges over the P basis functions numbered p = 1 to P. A basis set expansion may be a cosine series, a sine series, or more generally, a Fourier series.

For a one-dimensional data set, a preferred choice is a piecewise linear basis. The p^{th} basis function is zero outside the interval n_{p-1} to n_p, with n_0 taken as the left-most point of the data set and np as the right-most point. Within the interval,

$$s(n) = c_{p-1} + (c_p - c_{p-1})\,(n - n_{p-1})/(n_p - n_{p-1}).$$

For a one-dimensional or multi-dimensional data set, a preferred choice is a direct product of basis functions,

$$s(n) = \Sigma_p c_p \phi_p(n),$$

where here n and p are both d-dimensional and $\phi_p(n)$ can be expressed as

$$\phi_p(n) = \phi_{p1}(n_1)\,\phi_{p2}(n_2)\,\ldots\,\phi_{pd}(n_d)$$

where n_j and p_j are the components of n and p in dimension j and $\phi_{pj}(n_j)$ is a one-dimensional basis set in dimension j.

A preferred choice for the one-dimensional basis sets in the multidimensional direct product is an orthogonal basis. A preferred choice for an orthogonal basis is a trigonometric basis,

$$\phi_{pj}(n) = \cos[(pj-1)\pi(n-n_{j0})/(n-n_{j1})]\,,$$

where n_{j0} and n_{j1} are the left-most and right-most points in dimension j. In one dimension, for example, with points n = 0 to l corresponding to distances 0 to L, the p^{th} basis function is

$$\phi_p(x) = \cos[(p-1)\pi x/L].$$

An advantage of this basis set is that the low-order basis functions describe low-frequency variations. Typically, the low-frequency variation in the scaling is larger in amplitude than the high-frequency variation. Therefore truncating a trigonometric basis at low order still provides a good approximation of the scaling function from a complete basis (P approaches infinity). A preferable choice is to choose a value of P such that the low-frequency noise in the

data occurs on a length scale of L/P or longer. Preferably for differential display, L = 400 nt and the noise length scale is approximately 100 nt, so P ≈ 4 is preferable. It is this feature of a basis set such as the presently described basis set that contributes significantly to overcoming or eliminating the effects of low frequency noise.

Other acceptable basis sets include, for example, polynomials ($\phi_p(n) = nP$), special functions, and wavelets, and are well-known in the art See, *e.g.*, Press *et al.*, NUMERICAL RECIPES IN C, THE ART OF SCIENTIFIC COMPUTING, Second Edition, Cambridge Univ. Press, Cambridge UK, 1992, Chapters 5, 12 and 13.

The coefficients c_p are selected to minimize the distance Dist[A(n),s(n)a(n)] or maximize the similarity Sim[A(n),s(n)a(n)]. Methods to perform this optimization are well-known in the art. Preferable methods are conjugate direction minimization or conjugate gradient minimization, which use linear algebra to optimize the P basis set coefficients simultaneously. See, *e.g.*, Press *et al.*, NUMERICAL RECIPES IN C, THE ART OF SCIENTIFIC COMPUTING, Second Edition, Cambridge Univ. Press, Cambridge UK, 1992, Chapter 10.

For a piecewise linear basis, a preferable approximation that is faster computationally than a full minimization is to obtain c_p from an interval surrounding n_p, preferably the interval from n_{p-1} to n_{p+1}, by minimizing the distance Dist[A(n),c_pa(n)] or maximizing the similarity Sim[A(n),c_pa(n)].

Preferably for differential display, the number of piecewise linear basis functions is selected so that the low-frequency noise in the data occurs on a length scale of L/(0.3P) or longer. With L = 400 nt and a noise length scale approximately 100 nt, P ≈ 13 is preferable (interpolation points spaced every 35 nt).

Iterative Scaling

A group of data sets $\{a_i(n)\}$ can be brought into closer agreement with each other by first normalizing each data set, then generating an average A(n), then scaling each data set $a_i(n)$ to the average A(n), then repeating these steps. If desired, the average A(n) can be re-normalized after every iteration.

Iterations continue until a termination condition has been satisfied. A preferable termination condition is that A(n) has converged. This means that the distance Dist[A(n),A'(n)] between the value of A(n) after an iteration and its value A'(n) after the next iteration is smaller

than some threshold value. Alternatively, the scaling functions $s_i(n)$ for each of the slave data sets $a_i(n)$ can be checked for convergence. A second preferable termination condition is that a predetermined number of iterations has been reached.

It is possible to allow multiple termination conditions, with iterations ending after just one condition is satisfied.

Note that the square distance measure essentially calculates the standard deviation of the data sets. Thus, minimizing the square distance is essentially identical with performing scaling that minimizes the standard deviation of the scaled traces.

Iterative scaling may occur at a hierarchy of levels including experimental replicates, independent individuals, and groups. Recall that the data set corresponding to experimental replicate j of organism i of group A is $a_{ij}(n)$. Similarly, the data sets $b_{ij}(n)$ are obtained for group B, data sets $c_{ij}(n)$ for group C, and so forth for each of the groupings.

In one implementation of iterative scaling, data sets are normalized, scaled, and averaged within each organism, then within each group, and then between groups. One process is as follows:

- For each individual i in each group A, compute the average $A_i(n)$ by iterative scaling of the data sets $a_{ij}(n)$ as follows:
    - Each data set in $a_{ij}(n)$ is normalized.
    - Initialize $A_i(n)$ as the average of the experimental replicates.
    - Repeat the following steps until $A_i(n)$ has converged or the number of iterations has reached a threshold:
        - Optimize the scaling function $s_{ij}(n)$ to bring $a_{ij}(n)$ into best agreement with $A_i(n)$.
        - Compute the new average $A_i(n)$ from the scaled sets $s_{ij}(n)a_{ij}(n)$.
        - Optionally normalize the new $A_i(n)$.
    - Calculate the standard deviation $SD_i(n)$ as a measure of the experimental variance between scaled data sets.

- For each group A, compute the average A(n) by iterative scaling of the individual averages $A_i(n)$ as follows:
  - Initialize A(n) as the average of the individual averages $A_i(n)$.
  - Repeat the following steps until A(n) has converged or the number of iterations has reached a threshold:
    - Optimize the scaling function $s_i(n)$ to bring each $A_i(n)$ into best agreement with the group average A(n).
    - Compute the new average A(n) from the scaled individual averages $s_i(n)A_i(n)$.
    - Optionally, normalize the new A(n).
  - Calculate the standard deviation $SD_A(n)$ as a measure of the variance between scaled individual averages.
- For the final scaling of groups to each other, perform one of the following two operations:
  - Option 1: scale by computing the grand mean M(n) from all the group averages A(n), B(n), ... , as follows:
    - Initialize the grand mean M(n) as the average of A(n), B(n), ...
    - Repeat the following steps until M(n) has converged or the number of iterations has reached a threshold:
      - Optimize the scaling functions $s_A(n)$, $s_B(n)$, ... , that bring A(n), B(n), ... , into best agreement with M(n).
      - Compute the new M(n) from the scaled group averages $s_A(n)A(n)$, $s_B(n)B(n)$, ...
      - Optionally, normalize the new M(n)
    - Calculate the standard deviation SD(n) as a measure of the variance between scaled group averages.

- Option 2: select one of the groups R(n) as a reference and scale the remaining groups to R(n). Calculate the standard deviation SD(n) as a measure of the variance between the scaled group averages.

Using this process, the scaling terms must be back-propagated to compare averages other than the final, scaled group averages. For example, if scaled group averages are required, $s(n)A(n)$ is used. If scaled individual averages are required, then $s(n)s_i(n)A_i(n)$ is used. If scaled data sets are required, then $s(n)s_i(n)s_{ij}(n)a_{ij}(n)$ is used.

In a second implementation, intermediate averages are not required. This implementation requires that a weighting method be selected. With a preferred weighting method, each data set from individual i is preferably given a weight proportional to 1/(number of replicates from individual i) . This gives each individual equal weight and prevents an individual with many replicates from dominating the average. Other preferable methods include weighting each data set equally and weighting each data sets to give each group equal weight. If each group has equal weight, one method is to weight each replicate equally. Thus, each data set from group A is given a weight proportional to 1/(number of replicates from all the individuals belonging to group A). An alternate method is to weight each data set variably to give each individual within a group equal weight. Thus, each data set from individual i of group A is given a weight proportional to 1/[(number of individuals in group A)(number of replicates in individual i)].

After selecting a weighting method, apply the following algorithm:

- Initialize the grand mean M(n) by averaging all the data sets $a_{ij}(n)$ according to the selected weighting method..
- Repeat the following steps until M(n) has converged or the number of iterations reaches a threshold:
  - Optimize the scaling functions $s_{ij}(n)$ to bring each $a_{ij}(n)$ into best agreement with M(n).
  - Compute the new M(n) from the scaled data sets $s_{ij}(n)a_{ij}(n)$ using the selected weighting method.
  - Optionally, normalize the new M(n).

- Calculate the individual averages $A_i(n)$ by averaging the scaled replicates $s_{ij}(n)a_{ij}(n)$ belonging to individual i with weights according to the selected weighting method. Calculate the standard deviation $SD_i(n)$ within each individual.

- Calculate the group averages A(n), B(n), ... , by averaging the individual averages $A_i(n)$ according to the selected weighting method. Calculate the standard deviation $SD_A(n)$, $SD_B(n)$, ... , within each group.

For each of the iterative scaling steps, a preferable threshold for differential display data is 2 iterations.

Jiggling

One aspect of difference finding is comparing the heights of peaks in two data sets. In many data sets, the same peak may occur at different positions in different data sets. For example, a peak in one replicate of data set may occur at position n, while in a second data set it may occur at position n+1 or n-1 due to experimental variability. (See FIG. 1)

A jiggling algorithm identifies the peak height in a data set a(n) that corresponds to a given location n'. A preferred jiggling algorithm requires a parameter w_{jiggle}, which describes the width of the jiggling window.

The preferred algorithm starts at position n' and searches for the peak in a(n) closest to n' and within distance w_{jiggle}. The height of a(n) at this peak position is termed the jiggled height of a(n) at n'. If two peaks are within equal distance, the higher value is preferably taken as the jiggled height. If there is no peak within distance w_{jiggle}, then the height a(n') is the jiggled height.

For a one-dimensional data set, for example, the data range for the jiggling peak search is $n'-w_{jiggle}$ through $n'+w_{jiggle}$. If n' is a peak in a(n), then the value a(n') is the jiggled height of a(n) at n'. Otherwise the positions $n'\pm1$, $n'\pm2$, ..., $n'\pm w_{jiggle}$ are tested in turn for peaks; if a peak is found at location n'', then a(n'') is the jiggled height of a(n) at n'. If no peak is found, then a(n') is the jiggled height.

Less preferably, all of the peaks within distance w_{jiggle} of n' are examined and the maximum value is taken as the jiggled height of a(n) at n'. For a one-dimensional data set, all of the peaks in the window $n'-w_{jiggle}$, ... , $n'-1$, n', $n'+1$, ... , $n'+w_{jiggle}$ are examined and the

maximum value is taken as the jiggled height of a(n) at n'. If there is no peak in the interval, then a(n') is the jiggled height.

For differential display data, a preferable value is $w_{jiggle}\Delta x = 0.4$ nt.

Difference Finding

Difference finding identifies locations where at least one of the groups has a peak and its value is significantly different from the other groups. The group averages and individual averages produced by iterative scaling serve as inputs to difference finding.

It is preferable to employ an algorithm that uses jiggling to identify corresponding peaks in different data sets. This avoids spurious differences due to slight offsets in peak positions.

It is also preferable to employ an algorithm that identifies at most one difference from peaks that correspond. A preferable method employs a parameter w_{diff} that defines the minimum distance between differences. A preferable choice is $w_{diff} > w_{peak}$. For differential display data, a preferable value is $w_{diff}\Delta x = 1.1$ nt.

A preferred algorithm is as follows:

- Generate a master peak mask $M_{peak}(n)$ using one of the following alternatives:
  - Option 1: For each group A and individual A_i, calculate a peak mask $m_{peak}(n)$ from the individual average $A_i(n)$. Then, for each position n, $M_{peak}(n)$ is 1 if at least one of the individual peak masks is 1 and is 0 otherwise.
  - Option 2: Calculate the peak mask $M_{peak}(n)$ directly from the grand mean M(n) of all the groups.
  - Option 3: If there are only two groups A and B, generate a peak mask from the difference A(n) — B(n).
- For each position n that appears in the peak mask $M_{peak}(n)$, calculate the significance of a difference as follows:
  - For each individual i in each group A, find the jiggled height of A_i at n.
  - Calculate the group averages based on the jiggled heights.
  - Perform an F-test that compares the variance between group averages with the variance within groups.

- Associate the p-value of the F-test with the position n. Also record the number of samples that had a peak at position n.

- Generate a list of peak positions sorted from lowest p-value to highest p-value. If two positions have the same p-value, break the tie by listing first the position with more sample peaks. Break any remaining ties by listing first the lower position.

- Repeat the following steps until the list is empty:

  - Remove the first element from the list and record its position n as a difference.

  - Strike out any remaining elements in the list that are within distance w_{diff} of n. For a one-dimensional data set, for example, strike out any differences at positions $n\pm1, n\pm2, \ldots, n\pm w_{diff}$.

For difference finding with two groups A and B, a pooled variance t-test may be employed instead of an F-test.

A less preferred algorithm for a comparison between two groups, A and B, and one-dimensional data set, is as follows:

- Perform the final step of iterative scaling by scaling A(n) to B(n).

- Calculate the peak mask $M_{peak}(n)$ from the difference A(n) — B(n).

- Generate a list of peak positions sorted from smallest n to largest n.

- Initialize a variable LASTPOSITION as 0 and a variable LASTDIRECTION as 0.

- Repeat the following steps until the list of peaks is empty:

  - Remove the first element n from the list and calculate p-value(n) from a t-test comparing the jiggled heights of individuals from group A to the jiggled heights of individuals from group B at position n. If the average of the group A heights is larger than the average of the group B heights, then direction(n) = +1; otherwise, direction(n) = -1.

  - If direction(n) is not equal to LASTDIRECTION, or if (n — LASTPOSITION) > w_{diff}, then

    - If LASTPOSITION is not 0, save LASTPOSITION as a difference with p-value equal to LASTPVALUE

- Update LASTPOSITION = n, LASTDIRECTION = direction(n), and LASTPVALUE = p-value(n).

- Otherwise if p-value(n) < LASTPVALUE then

  - Update LASTPOSITION = n, LASTDIRECTION = direction(n), and LASTPVALUE = p-value(n).

- Otherwise

  - Continue with the next peak from the list.

- If LASTDIRECTION is not 0, then save the final difference at position LASTPOSITION with p-value equal to LASTPVALUE.

EXAMPLE

Example 1. Differential Gene Expression in Phenobarbitol-Treated Rats

Male Sprague-Dawley rats (Harlan Sprague Dawley, Inc., Indianapolis, Indiana) of 10-14 weeks of age were gavage-fed and dosed with phenobarbitol once a day for three days at a dose of 3.81 mg/kg/day. The drug was dissolved in sterile water prior to treatment. This dosage corresponds to the ED100 (the upper limit of the effective dose for humans) adjusted for the difference in metabolic rate between rats and humans. Three rats were used for the drug treatment group, and an additional three rats were treated with sterile water to serve as the control group.

Rats were sacrificed 24 hours after the final dose and their brains were harvested. Collection of mRNA from the harvested brains, synthesis of the corresponding cDNA, and differential display protocols were done as has been described elsewhere. See, U. S. Patent No. 5,871,697; Shimkets *et al. Nat. Biotechnol.* 17: 798-803 (1999).

Three experimental replicate raw traces were collected from each of two individual animals, one treated with phenobarbital and the other treated with sterile water. Data points were collected for fragments from 30 nt to 450 nt in length, and the data sets were discretized for a point every 0.1 nt. The averages of the 3 replicate raw traces for each individual, prior to normalization or scaling, are shown in FIG. 4. Each trace was weighted equally, and no data were masked for either noise or saturation.

Next, noise was masked using $I_{noise} = 500$ and $w_{sat} = 5$, and peaks were identified in each replicate using $w_{peak} = 3$ with the condition that all 7 points contributing to a peak be above the noise level and not saturated. The peaks were sorted in increasing order of height, and each replicate was normalized to give the 75th percentile peak an intensity of 100. The 3 normalized traces were averaged for each individual, and the individual averages are displayed in FIG. 5.

For all the scaling operations that follow, the basis set used was piecewise linear with 13 scaling points located every 35 nt beginning at 30 nt and ending at 450 nt. The distance function was $[\ln(a(n)/A(n))]^2$, and points for which $|\ln(a(n)/A(n)]| > 3$ were masked out.

The first step in the scaling procedure was that the 3 normalized traces for each individual were averaged, scaled to the average, re-averaged, then re-scaled to the average for 2 rounds of iterative scaling. Next, the phenobarbital-treated individual average and the sterile-water-treated individual average were themselves averaged, the individual averages scaled to the grand average, then the process repeated for 2 rounds of iterative scaling. Finally, the phenobarbital-treated individual average was scaled to the sterile-water-treated individual average. The final scaling factors are shown in FIG. 6 for the two individuals. The final individual averages are shown in FIG. 7.

With both the normalized traces and the normalized and scaled traces, difference finding was performed using $w_{peak} = 2$, $w_{jiggle} = 4$, and $w_{diff} = 11$. Significance levels were calculated as 1 – p-value from a t-test based on the scaled replicate traces. See Table 1, below. Only differences with a significance greater than 0.9 and a ratio $|\ln(\text{Phenobarbital/Water})| > \ln(1.5)$ were retained. The differences identified are listed in the table below (significances greater that 0.99 are reported as 1). The normalized traces generated 35 differences, whereas the scaled traces generated 18 differences, of which 12 were in common with the normalized traces. The differences in the normalized traces that are removed by scaling tend to have lower significance than the differences that are retained, indicating that scaling helps identify the differences with greater support in the data..

Table 1

Length	t-test (1 - p-value)	
	Normalized Only	Normalized and Scaled
51.9	1	1
53.6	1	1
67.3	0.94	-
81.8	0.99	0.97
87.1	-	1
103.2	-	1
120.8	1	0.95
127.5	1	-
154.6	-	1
158.9	-	1
165.0	0.98	-
170.0	1	0.99
190.1	0.97	-
205.7	1	1
218.6	1	0.91
228.1	1	-
233.0	0.98	-
263.1	0.94	-
274.1	1	1
280.0	0.91	-
303.2	-	0.99
331.8	1	1
340.6	0.98	0.9
352.2	-	1
353.7	0.99	-
354.9	0.96	-
388.8	1	1
394.5	1	-
395.7	1	0.99
402.4	0.96	-
404.7	1	-
406.4	1	-
437.7	0.99	-
443.1	0.98	-
447.9	1	-

EQUIVALENTS

From the foregoing detailed description of the specific embodiments of the invention, it should be apparent that unique methods for identifying differences between at least two data sets have been described. Although particular embodiments have been disclosed herein in detail, this has been done by way of example for purposes of illustration only, and is not intended to be limiting with respect to the scope of the appended claims which follow. In particular, it is

contemplated by the inventor that various substitutions, alterations, and modifications may be made to the invention without departing from the spirit and scope of the invention as defined by the claims. For instance, the choice of algorithms used to transform data sets, such as normalization calculations, averaging calculations or scaling calculations, or the choice of data sets to be analyzed is believed to be a matter of routine for a person of ordinary skill in the art with knowledge of the embodiments described herein.

CLAIMS

1. A method of identifying a difference between at least two groups, the method comprising:

a) providing a first group having one or more elements in a first data set;

b) applying at least one transformation to said first data set to provide a transformed data set, wherein said transformation is a calculation selected from a normalizing calculation, an averaging calculation and a scaling calculation; and

c) distinguishing differences, if present, between elements of said first transformed data set and a second groups having one or more elements in a second data set;

thereby identifying a difference between the groups.

2. The method of claim 1, wherein said second data set is a transformed data set.

3. The method of claim 1, wherein said transformation minimizes noise associated with one or more signals associated with elements of at least one data set.

4. The method of claim 3, wherein said noise comprises low frequency noise.

5. The method of claim 3, wherein said noise comprises jiggle.

6. The method of claim 5, wherein said jiggle includes positional shifts of corresponding elements between said first and second data sets.

7. The method of claim 1, wherein elements of at least one data set comprise a trace.

8. The method of claim 7, wherein said trace represents the result of an electrophoretogram or a chromatogram.

9. The method of claim 1, wherein elements of at least one data set comprise one or more positions in an array.

10. The method of claim 9, wherein any one position in the array is used to determine an extent of matching between a reagent affixed to the array at said position and a sample contacting the array position.

11. The method of claim 10, wherein the reagent is a first nucleic acid and the sample comprises a second nucleic acid, wherein the second nucleic acid is in a hybridization solution mixture, and the matching constitutes hybridization of a sample nucleic acid to the affixed nucleic acid.

12. The method of claim 1, wherein at least one data set comprises elements derived from an analysis of one or more differentially expressed nucleic acids.

13. The method of claim 1, wherein at least one data set is derived from an analysis of one or more individuals in a group.

14. The method of claim 13, wherein at least one data set is subjected to a masking operation.

15. The method of claim 13, wherein the elements in the data sets comprise values derived from an analysis of differentially expressed nucleic acids in the plurality of individuals in a group.

16. The method of claim 15, wherein a data set obtained for each group is obtained by applying at least one transformation to the data set from each individual, wherein said transformation is selected from the group consisting of a normalizing calculation, an averaging calculation and a scaling calculation.

17. The method of claim 15, wherein each individual provides at least one replicate sample.

18. The method of claim 15, wherein the data set from any one individual or replicate comprises a data set derived from a trace of an electropherogram or a chromatogram.

19. The method of claim 18, wherein said data set is transformed by applying at least one of a normalizing calculation, an averaging calculation and a scaling calculation to the data set from each individual or replicate.

20. The method of claim 18, wherein said data set is discretized prior to the normalizing, the scaling and/or the averaging calculation.

21. The method of claim 1, wherein said transformation comprises a normalization calculation.

22. The method of claim 21, wherein said normalization calculation comprises adjusting each data set such that a subset of elements in each data set has similar or identical values.

23. The method of claim 21, wherein at least one data set is subjected to a masking operation.

24. The method of claim 1, wherein said transformation comprises an averaging calculation.

25. The method of claim 24, wherein said averaging calculation comprises calculating an average for a location or for a discretized position of said first and second data sets, and wherein the average may be an unweighted average or a weighted average.

26. The method of claim 1, wherein said transformation comprises a scaling calculation.

27. The method of claim 26, wherein said scaling calculation comprises a calculation that causes a first data set to resemble a second data set, provided that an element in the scaled first data set whose intensity differs significantly from the intensity of the element in the second data set at the same location or the same position contributes to identifying the difference between the data sets.

28. The method of claim 27, wherein scaling comprises calculating a scaling function based on optimization of a distance between the data sets.

29. The method of claim 27, wherein scaling comprises calculating a scaling function based on optimization of a similarity between the data sets.

30. The method of claim 27, wherein the scaling calculation employs a scaling function.

31. The method of claim 30, wherein the scaling function is a basis set expansion.

32. The method of claim 31, wherein the basis set is a piecewise linear basis set.

33. The method of claim 31, wherein the basis set is a Fourier series.

34. The method of claim 30, wherein the scaling function is a direct product of basis functions.

35. The method of claim 1, wherein the transformation comprises successive iterations of a cycle of calculations that are carried out until a specified termination condition has been satisfied, wherein the calculations comprise at least one of a normalization calculation, an averaging calculation and a scaling calculation.

36. The method of claim 35, wherein the termination condition is met when a transformed data set has converged.

37. The method of claim 35, wherein the termination condition is met when a predetermined number of iterations of a cycle of calculations has been reached.

38. The method of claim 1, wherein elements within any two or more data sets have been corrected for signal alignment.

39. The method of claim 1, wherein the distinguishing of differences between the elements of the resulting data sets comprises an application of at least one difference finding algorithm.

40. A display means displaying a representation of a difference between two or more transformed data sets, wherein each data set comprises ordered elements; and the data sets are transformed by at least one calculation selected from the group consisting of a normalizing calculation, an averaging calculation and a scaling calculation.

41. The display means of claim 40, wherein the representation is obtained by a process comprising the steps of:

a) providing a first group having one or more elements in a first data set;

b) applying at least one transformation to said first data set to provide a transformed data set, wherein said transformation is a calculation selected from a normalizing calculation, an averaging calculation and a scaling calculation; and

c) distinguishing differences, if present, between elements of said first transformed data set and a second groups having one or more elements in a second data set;

thereby identifying a difference between the groups.

42. The display means of claim 41, wherein said second data set is a transformed data set.

43. The display means of claim 41, wherein said transformation minimizes noise associated with one or more signals associated with elements of at least one data set.

44. The display means of claim 43, wherein said noise comprises low frequency noise.

45. The display means of claim 43, wherein said noise comprises jiggle.

46. The display means of claim 45, wherein said jiggle includes positional shifts of corresponding elements between said first and second data sets.

47. The display means of claim 41, wherein elements of at least one data set comprise a trace.

48. The display means of claim 47, wherein said trace represents the result of an electrophoretogram or a chromatogram.

49. The display means of claim 41, wherein elements of at least one data set comprise one or more positions in an array.

50. The display means of claim 49, wherein any one position in the array is used to determine an extent of matching between a reagent affixed to the array at said position and a sample contacting the array position.

51. The display means of claim 50, wherein the reagent is a first nucleic acid and the sample comprises a second nucleic acid, wherein the second nucleic acid is in a hybridization solution mixture, and the matching constitutes hybridization of a sample nucleic acid to the affixed nucleic acid.

52. The display means of claim 41, wherein at least one data set comprises elements derived from an analysis of one or more differentially expressed nucleic acids.

53. The display means of claim 41, wherein at least one data set is derived from an analysis of one or more individuals in a group.

54. The display means of claim 53, wherein at least one data set is subjected to a masking operation.

55. The display means of claim 53, wherein the elements in the data sets comprise values derived from an analysis of differentially expressed nucleic acids in the plurality of individuals in a group.

56. The display means of claim 55, wherein a data set obtained for each group is obtained by applying at least one transformation to the data set from each individual, wherein said transformation is selected from the group consisting of a normalizing calculation, an averaging calculation and a scaling calculation.

57. The display means of claim 55, wherein each individual provides at least one replicate sample.

58. The display means of claim 55, wherein the data set from any one individual or replicate comprises a data set derived from a trace of an electropherogram or a chromatogram.

59. The display means of claim 58, wherein said data set is transformed by applying at least one of a normalizing calculation, an averaging calculation and a scaling calculation to the data set from each individual or replicate.

60. The display means of claim 58, wherein said data set is discretized prior to the normalizing, the scaling and/or the averaging calculation.

61. The display means of claim 41, wherein said transformation comprises a normalization calculation.

62. The display means of claim 61, wherein said normalization calculation comprises adjusting each data set such that a subset of elements in each data set has similar or identical values.

63. The display means of claim 61, wherein at least one data set is subjected to a masking operation.

64. The display means of claim 41, wherein said transformation comprises an averaging calculation.

65. The display means of claim 64, wherein said averaging calculation comprises calculating an average for a location or for a discretized position of said first and second data sets, and wherein the average may be an unweighted average or a weighted average.

66. The display means of claim 41, wherein said transformation comprises a scaling calculation.

67. The display means of claim 66, wherein said scaling calculation comprises a calculation that causes a first data set to resemble a second data set, provided that an element in the scaled first data set whose intensity differs significantly from the intensity of the element in the second data set at the same location or the same position contributes to identifying the difference between the data sets.

68. The display means of claim 67, wherein scaling comprises calculating a scaling function based on optimization of a distance between the data sets.

69. The display means of claim 67, wherein scaling comprises calculating a scaling function based on optimization of a similarity between the data sets.

70. The display means of claim 67, wherein the scaling calculation employs a scaling function.

71. The display means of claim 70, wherein the scaling function is a basis set expansion.

72. The display means of claim 71, wherein the basis set is a piecewise linear basis set.

73. The display means of claim 71, wherein the basis set is a Fourier series.

74. The display means of claim 70, wherein the scaling function is a direct product of basis functions.

75. The display means of claim 41, wherein the transformation comprises successive iterations of a cycle of calculations that are carried out until a specified termination condition has been satisfied, wherein the calculations comprise at least one of a normalization calculation, an averaging calculation and a scaling calculation.

76. The display means of claim 75, wherein the termination condition is met when a transformed data set has converged.

77. The display means of claim 75, wherein the termination condition is met when a predetermined number of iterations of a cycle of calculations has been reached.

78. The display means of claim 41, wherein elements within any two or more data sets have been corrected for signal alignment.

79. The display means of claim 41, wherein the distinguishing of differences between the elements of the resulting data sets comprises an application of at least one difference finding algorithm.

80. A representation of a difference between normalized, averaged and scaled data sets, wherein each data set comprises ordered elements.

81. The representation of claim 80 wherein the representation is obtained by a process comprising the steps of

a) providing a first group having one or more elements in a first data set;

b) applying at least one transformation to said first data set to provide a transformed data set, wherein said transformation is a calculation selected from a normalizing calculation, an averaging calculation and a scaling calculation; and

c) distinguishing differences, if present, between elements of said first transformed data set and a second groups having one or more elements in a second data set;

thereby identifying a difference between the groups.

82. The representation of claim 81, wherein said second data set is a transformed data set.

83. The representation of claim 81, wherein said transformation minimizes noise associated with one or more signals associated with elements of at least one data set.

84. The representation of claim 83, wherein said noise comprises low frequency noise.

85. The representation of claim 83, wherein said noise comprises jiggle.

86. The representation of claim 85, wherein said jiggle includes positional shifts of corresponding elements between said first and second data sets.

87. The representation of claim 81, wherein elements of at least one data set comprise a trace.

88. The representation of claim 87, wherein said trace represents the result of an electrophoretogram or a chromatogram.

89. The representation of claim 81, wherein elements of at least one data set comprise one or more positions in an array.

90. The representation of claim 89, wherein any one position in the array is used to determine an extent of matching between a reagent affixed to the array at said position and a sample contacting the array position.

91. The representation of claim 90, wherein the reagent is a first nucleic acid and the sample comprises a second nucleic acid, wherein the second nucleic acid is in a hybridization solution mixture, and the matching constitutes hybridization of a sample nucleic acid to the affixed nucleic acid.

92. The representation of claim 81, wherein at least one data set comprises elements derived from an analysis of one or more differentially expressed nucleic acids.

93. The representation of claim 81, wherein at least one data set is derived from an analysis of one or more individuals in a group.

94. The representation of claim 93, wherein at least one data set is subjected to a masking operation.

95. The representation of claim 93, wherein the elements in the data sets comprise values derived from an analysis of differentially expressed nucleic acids in the plurality of individuals in a group.

96. The representation of claim 95, wherein a data set obtained for each group is obtained by applying at least one transformation to the data set from each individual, wherein said transformation is selected from the group consisting of a normalizing calculation, an averaging calculation and a scaling calculation.

97. The representation of claim 95, wherein each individual provides at least one replicate sample.

98. The representation of claim 95, wherein the data set from any one individual or replicate comprises a data set derived from a trace of an electropherogram or a chromatogram.

99. The representation of claim 98, wherein said data set is transformed by applying at least one of a normalizing calculation, an averaging calculation and a scaling calculation to the data set from each individual or replicate.

100. The representation of claim 98, wherein said data set is discretized prior to the normalizing, the scaling and/or the averaging calculation.

101. The representation of claim 81, wherein said transformation comprises a normalization calculation.

102. The representation of claim 101, wherein said normalization calculation comprises adjusting each data set such that a subset of elements in each data set has similar or identical values.

103. The representation of claim 101, wherein at least one data set is subjected to a masking operation.

104. The representation of claim 81, wherein said transformation comprises an averaging calculation.

105. The representation of claim 104, wherein said averaging calculation comprises calculating an average for a location or for a discretized position of said first and second data sets, and wherein the average may be an unweighted average or a weighted average.

106. The representation of claim 81, wherein said transformation comprises a scaling calculation.

107. The representation of claim 106, wherein said scaling calculation comprises a calculation that causes a first data set to resemble a second data set, provided that an element in the scaled first data set whose intensity differs significantly from the intensity of the element in

the second data set at the same location or the same position contributes to identifying the difference between the data sets.

108. The representation of claim 107, wherein scaling comprises calculating a scaling function based on optimization of a distance between the data sets.

109. The representation of claim 107, wherein scaling comprises calculating a scaling function based on optimization of a similarity between the data sets.

110. The representation of claim 107, wherein the scaling calculation employs a scaling function.

111. The representation of claim 110, wherein the scaling function is a basis set expansion.

112. The representation of claim 111, wherein the basis set is a piecewise linear basis set.

113. The display means of claim 111, wherein the basis set is a Fourier series.

114. The representation of claim 110, wherein the scaling function is a direct product of basis functions.

115. The representation of claim 81, wherein the transformation comprises successive iterations of a cycle of calculations that are carried out until a specified termination condition has been satisfied, wherein the calculations comprise at least one of a normalization calculation, an averaging calculation and a scaling calculation.

116. The representation of claim 115, wherein the termination condition is met when a transformed data set has converged.

117. The representation of claim 115, wherein the termination condition is met when a predetermined number of iterations of a cycle of calculations has been reached.

118. The representation of claim 81, wherein elements within any two or more data sets have been corrected for signal alignment.

119. The representation of claim 81, wherein the distinguishing of differences between the elements of the resulting data sets comprises an application of at least one difference finding algorithm.

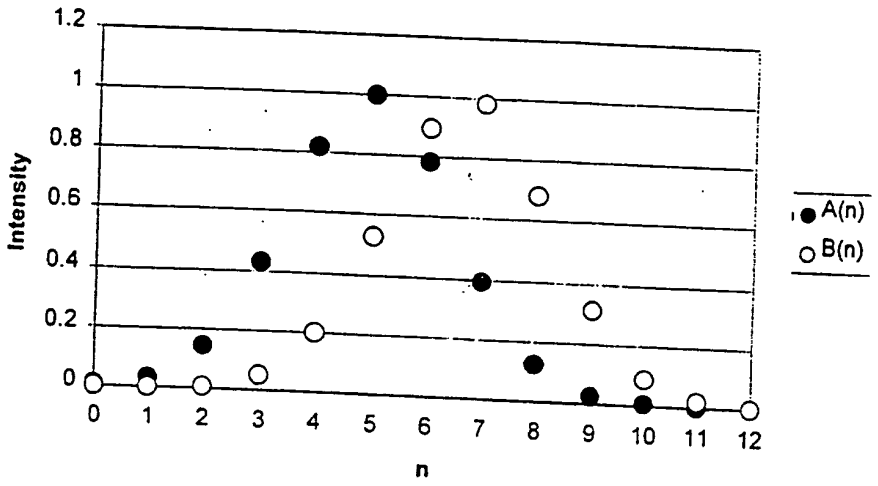


FIGURE 1

SHEET 2 OF 7

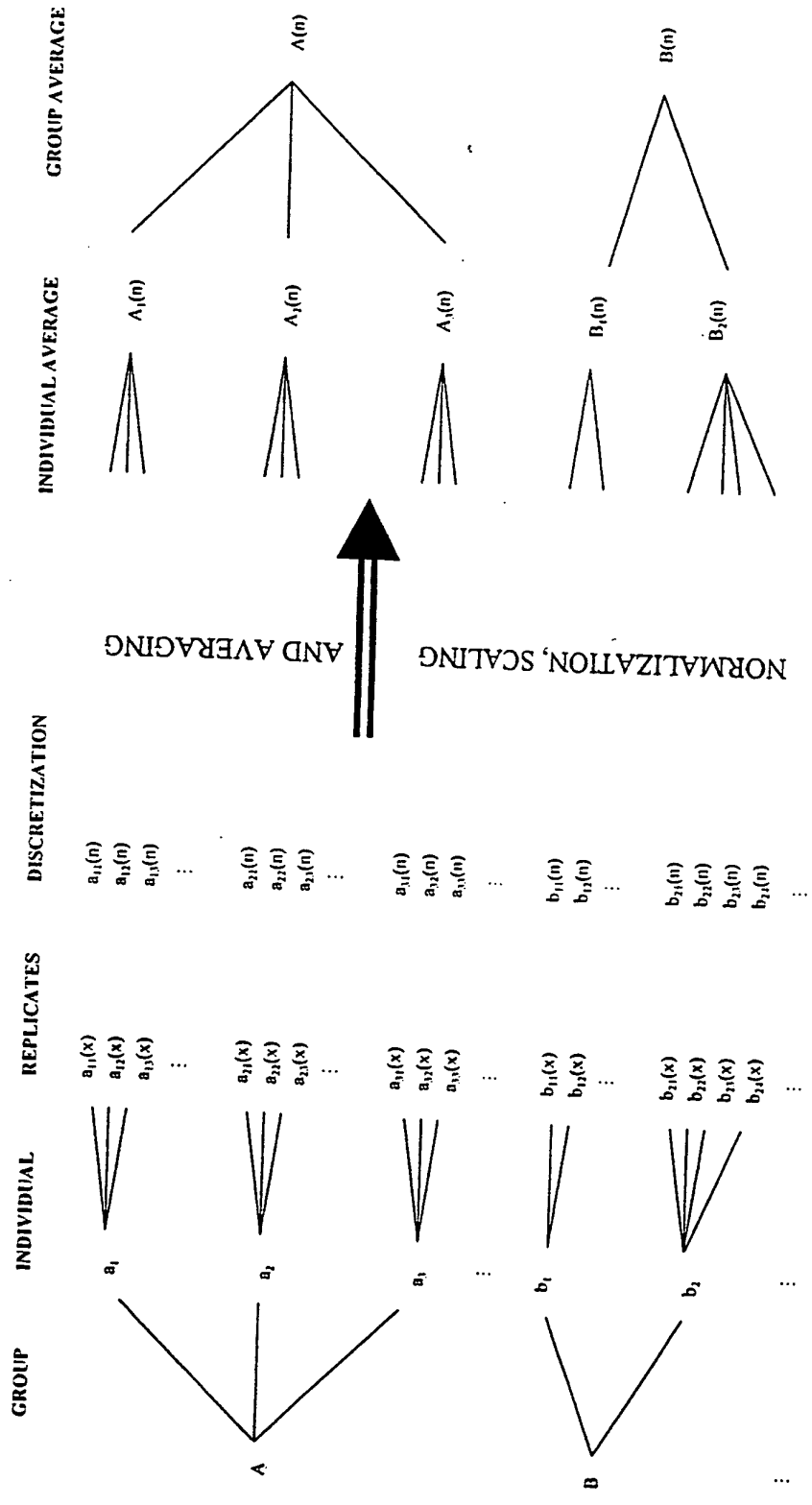


FIGURE 2

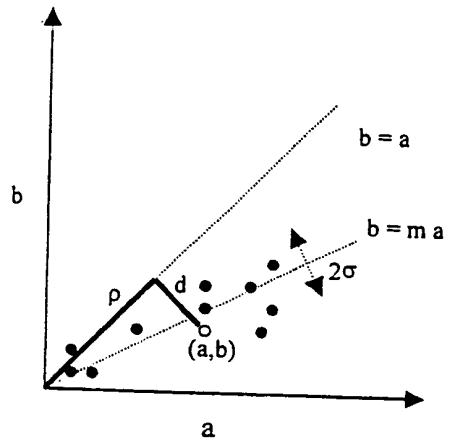


FIGURE 3

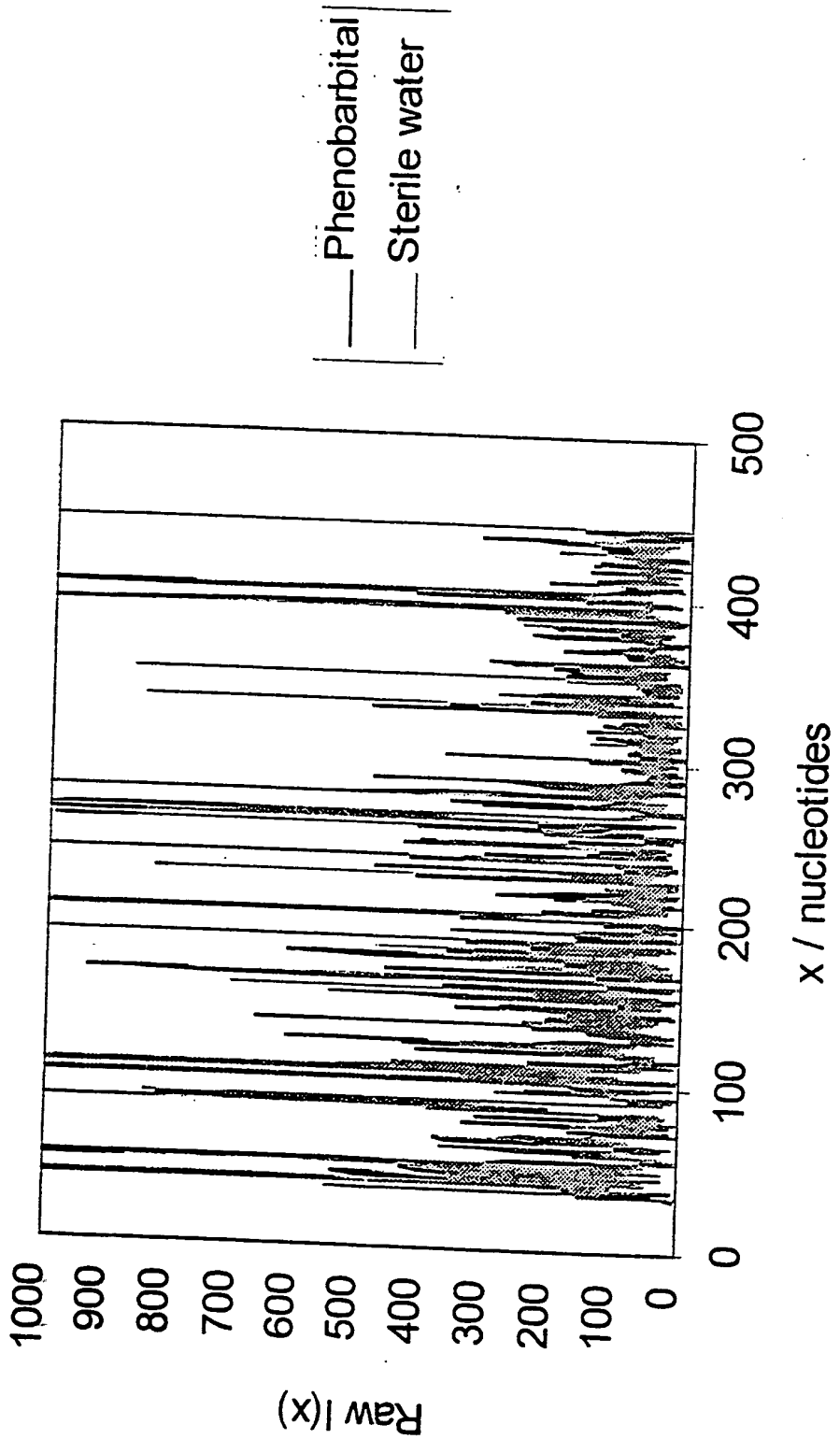

Phenobarbital
Sterile water
Raw I(x)
1000
900
800
700
600
500
400
300
200
100
0
0
100
200
300
400
500
x / nucleotides

FIGURE 4

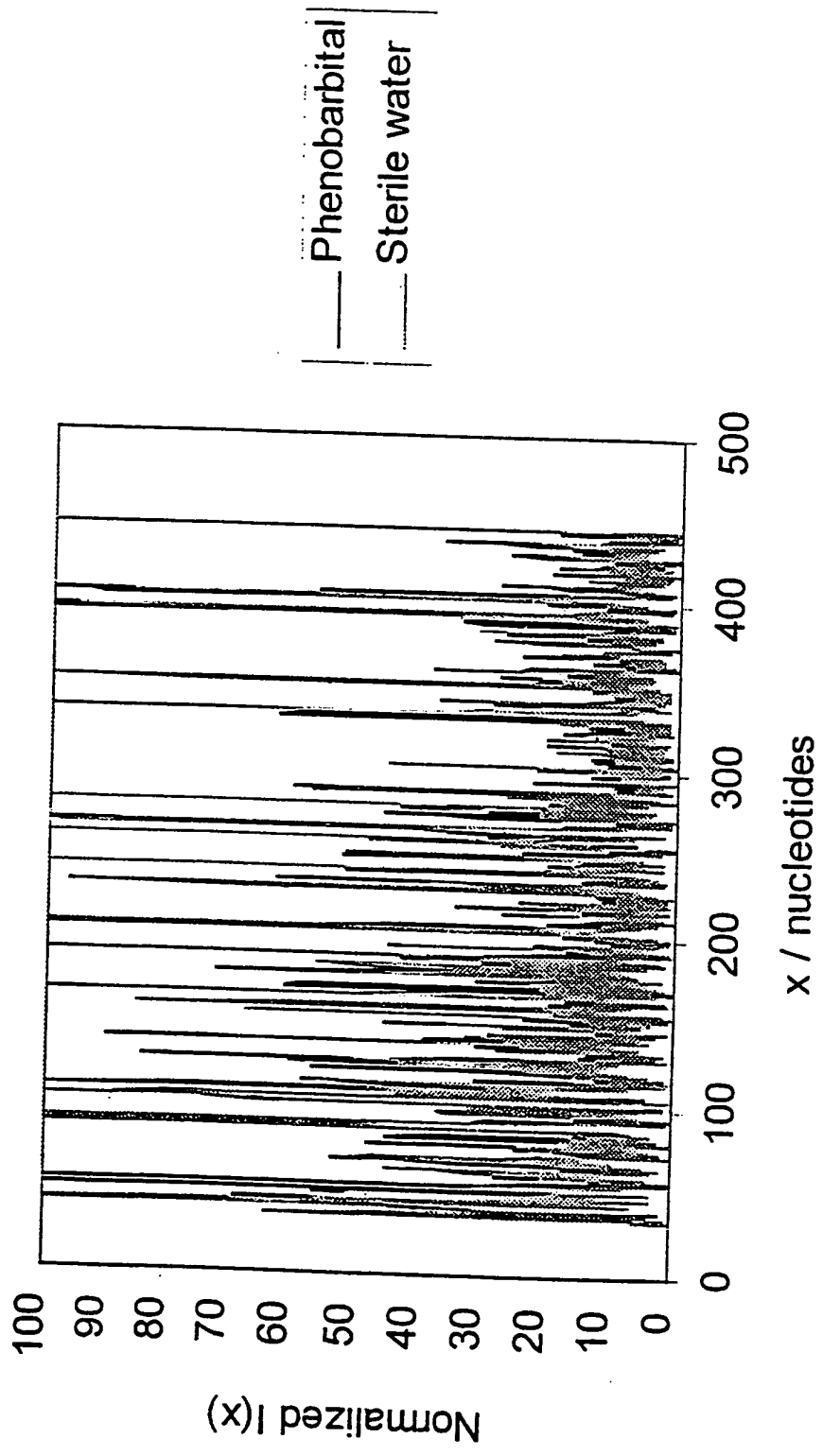

Phenobarbital
Sterile water
Normalized I(x)
100
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
x / nucleotides

FIGURE 5

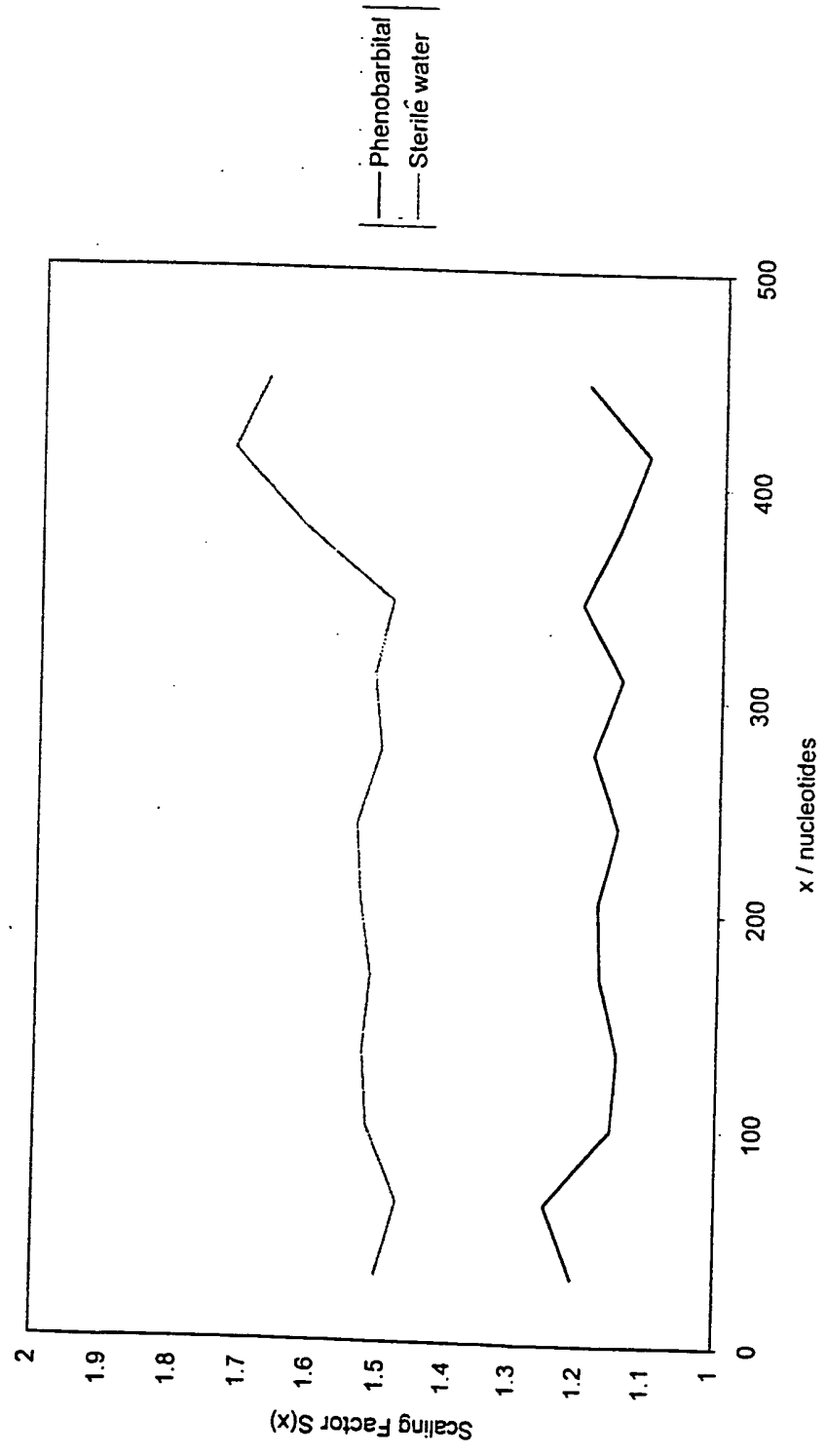

Phenobarbital
Sterile water
Scaling Factor S(x)
x / nucleotides
2
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
0
100
200
300
400
500

FIGURE 6

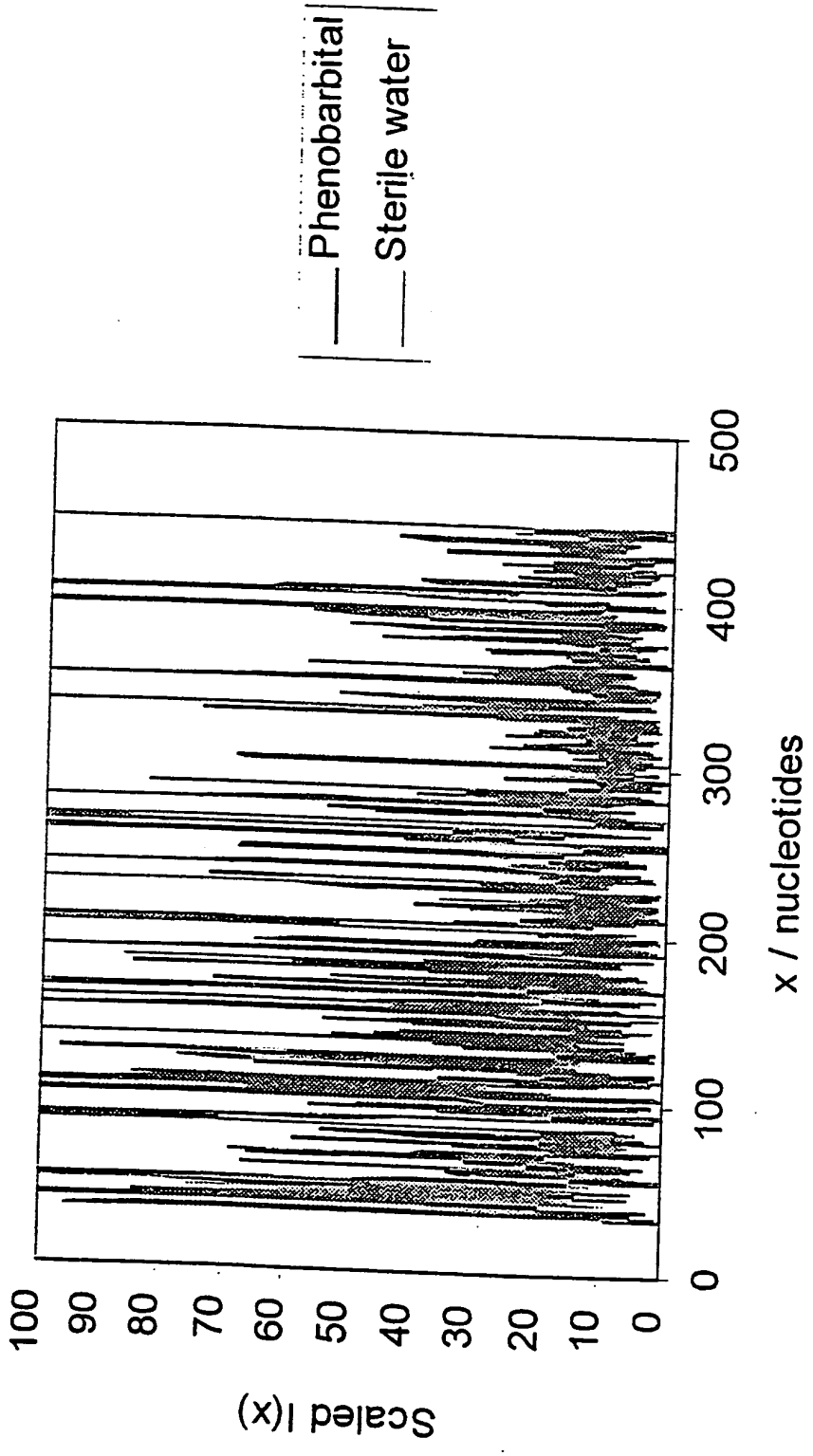

Phenobarbital
Sterile water
Scaled I(x)
100
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
x / nucleotides

FIGURE 7